ETL Report

Jin Hee Lee
03/30/2023

Introduction

Two datasets were introduced: Behavioral Risk Factor Data for Health-Related Quality of Life as presented by the CDC, and Us population by zip code. The primary data is Health-Related Quality of Life. Both datasets are big, with 1,622,831 and 126,464 rows included. The Health-Related Quality of Life dataset has several columns with only one value, no value, or null values. The column names also need to be changed to make the points of the values clear.

Data Sources

Behavioral risk factor HRQOL - dataset by CDC. data.world. (2017, February 2). Retrieved March 28, 2023, from https://data.world/cdc/behavioral-risk-factor-hrqol

Mvalcic. (2018, January 20). *Add city, state, longitude, and latitude data*. Kaggle. Retrieved March 28, 2023, from https://www.kaggle.com/code/mvalcic/add-city-state-longitude-and-latitude-data

Extraction

I found these datasets the following websites through Teachable:

https://dev-10.teachable.com/courses/data-engineering-c50/lectures/45590983, and both datafiles were extracted as CSV files from these websites. To download the files, open a blank Exel worksheet and use the ‘Get Data’ drop-down menu on the ‘Data’ tab. Select ‘From File’ and then choose ‘From Excel Workbook.’ This will open the folder, choose the file and click ‘Import’.

Transformation

1. After the files are uploaded, click the ‘Transform Data’ button, which will lead to the Power Query.

2. Remove the ‘category’, ‘CategoryId’, ‘Data source’, ‘Data value unit’, ‘Datavalue footnote_symbol’, ‘datavalue Std_Err’ columns as they have no value or only one value in them.
3. Check and clear rows with blanks and null values.
4. Change the column name ‘LocationAbbr’ to ‘State_abbr; ’location Desc’ to ‘State_name’.
5. Split the Geolocation column and name those latitude and longitude.
6. Leave only the following columns: Year, State_abbr, Satet, Topic, Question, Data_Value, Sample_Size, Break_Out, and Break_Out_Category.
7. Name above table as ‘TempSurveyqol’
8. remove the rest of the columns.
9. The ‘TempSurveyqol” table should have 101,429 lows and 10 columns.
10. When using the data in a table, watch out for value redundancy that can cause the total number of values to be double or triple.
11. Save the transformed table as a CSV files.
12. Make an ERD form a denomoralized table :’TempSurveyqol’
13. For ERD, make three referenced tables named ‘Location’, ‘Question’, ‘Breakout’, and the primary table named ‘Survey’
14. Make sure put all the appropriate data types for each column.
15. Create a data base named ‘Aqol’
16. Create a normalized schema in a Aqol database named ‘NormalizedSurvey’.
17. Make DDL for the four normalized tables.

Load

1. Export the denormalized file ‘TempSurveyqol’ into Aqol database using the import wizard.
2. Set appropriate data types for each column and decide if it is nullable.
3. There will not be null values because all the blanks and null values were cleared.
4. Look for the error messages and adjust using previous button.
5. Keep trying to import Data until it succeeds.
6. Import ‘USpopulation by zipcode’ data into ‘aqol’ database by using same steps as TempSurveyqol denomalized table.
7. Make 4 tables named ‘Survey’, ‘Breakout, ‘Question’, and ‘Location’ using DDL, and insert values on each table using inset query.
8. Make sure to put constraint on the primary table(‘Survey’) for reference tables using foreign keys.

9. After making four table with values in, then analyzed the relation with Power Bi and pivot table.

Conclusion

The above ETL (Extract, Transform, Load) process allows me to transform the big data into a format that can be easily analyzed or queried. It is particularly important when I was dealing with large databases such as Behavioral Risk Factor Data for Health-Related Quality of Life, and Us population by zip code. ETL process helped me to reduce the processing time and ensure the data quality by cleaning and filtering out irrelevant or erroneous data. It also helped to prevent data anomalies or inconsistencies, which might be leading to inaccurate or incomplete analysis. In addition, this ETL process helped me to optimize data storage by reducing redundancy and organizing data in a way that is efficient and easy to access. This can be particularly important when dealing with big data, as it can help to reduce storage costs and make it easier to scale the data infrastructure.